UNCLASSIFIED

AD 260 606

Reproduced
by the

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

CATALOGED BY ASTIA
AS AD No. 260606

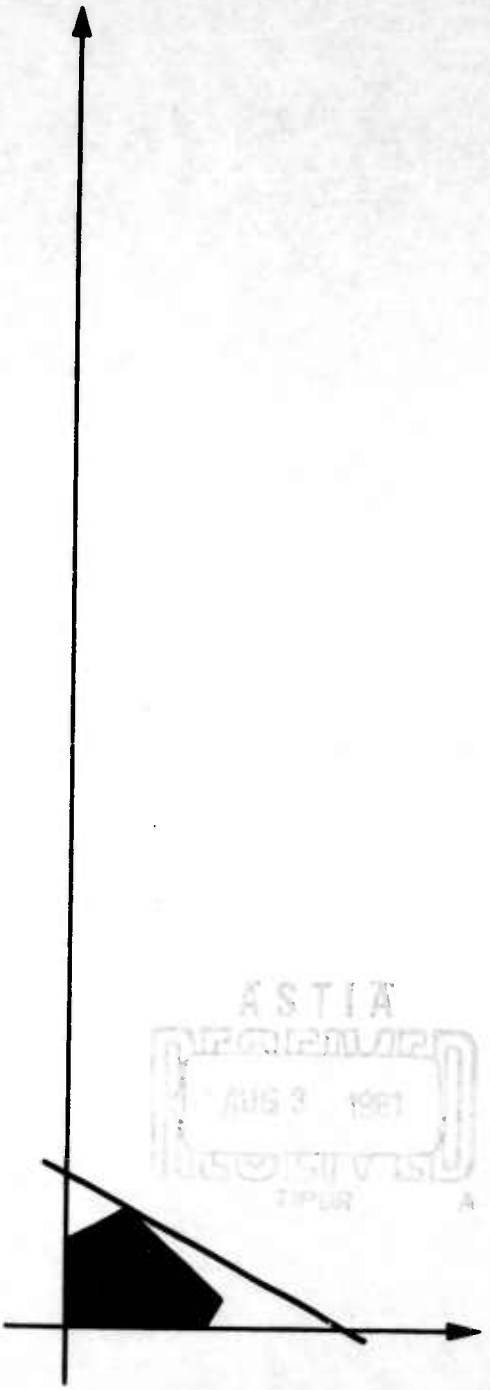
RESEARCH REPORT 1
3 MARCH 1961
I.E.R. 172-1

A FILING PROBLEM

by

William S. Jewell

XEROX 1-4-1



ASTIA
AUG 3 1961

OPERATIONS RESEARCH CENTER

INSTITUTE OF ENGINEERING RESEARCH

UNIVERSITY OF CALIFORNIA - BERKELEY

A FILING PROBLEM

by

William S. Jewell
Operations Research Center
University of California, Berkeley

This research was supported in part by the Office of Naval Research under contract Nonr-222(83) with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

March 3, 1961

Research Report 1

From a known distribution, $F(.)$, $\underline{n}$ samples are drawn, and are ordered as $b_1, b_2, \ldots, b_n$, according to the value of b ($0 \leqslant b \leqslant \infty$). For example, the b's might be family names, drawn from some known ethnic distribution, and arranged alphabetically for filing purposes. These ordered samples are to be divided into groups, and stored in filing drawers (or blocks of computer storage), with space left in each drawer for additions to the file. If many additions are made to the file, of course, there is the possibility that a drawer will overflow, and that the filing groups will have to be reorganized. Given a fixed initial sample size and common capacity of each drawer, the best system design will group the names so that the overflow probability is the same for each drawer. In this paper, we shall indicate how to perform this initial a priori grouping of the files.

Suppose that a certain drawer begins with the t^{th} sample whose value is b_t, and ends with the $(t+d)^{th}$ sample whose value is b_{t+d}. Now suppose that $\underline{N}$ new independent samples are drawn from the same distribution, and added to the files in correct order. The probability that $\underline{x}$ of the new samples will be placed in this particular drawer (i.e., that $\underline{x}$ of their values lie between b_t and b_{t+d}) is just:

$$P_x(N,n \mid b_t, b_{t+d}) = \binom{N}{x} p^x (1-p)^{N-x}, \quad \begin{matrix} d=0,1,..n \\ t=0,1,..n+1-d \\ x=0,1,..N \end{matrix} \qquad (1)$$

where $p = F(b_{t+d}) - F(b_t)$ is the probability that a single new sample would fall in this drawer, and where for convenience we have defined $b_0 = 0$ and $b_{n+1} = \infty$.

Usually, however, when a filing system is planned, it is not possible to know in advance the values of the first and last names in drawer, but only their file positions, the indices $\underline{t}$ and $\underline{t+d}$. Thus, one is interested in the probability

of insertation in a given drawer, averaged over all possible values, b_t and b_{t+d}, of the positions, t and t+d. It is known* that the joint distribution density of the t^{th} and $(t+a)^{th}$ ordered samples is:

$$\psi(b_t, b_{t+d}) = \frac{n!}{(t-1)!\,(d-1)!\,(n-t-d)!}\,[F(b_t)]^{t-1}\,\cdot \tag{2}$$

$$\cdot\,[F(b_{t+d}) - F(b_t)]^{d-1}\,[1 - F(b_{t+d})]^{n-t-d}\,f(b_t)\,f(b_{t+d})$$

so that the averaged distribution of new samples lying between the t^{th} and $(t+d)^{th}$ positions of the old samples is just:

$$w_x = w_x(N,n;t,t+d) = \int\int_{0 \le b_t \le b_{t+d} \le \infty} db_t\, db_{t+d}\, \psi(b_t, b_{t+d})\, P_x(N,n \mid b_t, b_{t+d}). \tag{3}$$

A first change of variable, $z = 1 - F(b_t)/F(b_{t+d})$, with respect to b_t; then a second change of variable, $y = F(b_{t+d})$; and we are led to the simplified form:

$$w_x = K \int_0^1 dy\, y^{x+t+d-1}\,(1-y)^{n-t-d} \int_0^1 dz\, z^{x+d-1}\,(1-z)^{t-1}\,(1-yz)^{N-x}, \tag{4}$$

$$\text{with} \quad K = (n!)\,[(t-1)!\,(d-1)!\,(n-t-d)!]^{-1}.$$

Upon expanding $(1-yz)^{N-x}$ with the binomial theorem, integrating the separated integrals using the definition of Beta functions, and summing the resulting series through the formula:

---

* See, for example, Gumbal, E. J., Statistics of Extremes, Columbia University Press, New York (1958), p. 53.

$$\sum_{k=0}^{m} \binom{m}{k} \frac{(a+k)!}{(a+b+k)!} (-)^k = \frac{a!\,(b+m-1)!}{(a+b+m)!\,(b-1)!}, \qquad (5)$$

one finds after some algebraic manipulation that:

$$w_x(N,n;t,t+d) = \frac{\binom{N+n-d-x}{n-d}\binom{d+x-1}{d-1}}{\binom{N+n}{n}}, \quad x = 0,1\ldots N \qquad (6)$$

the so-called distribution of exceedances, whose properties are known*. For example, the mean number of names (out of the N new samples) which lie between the t^{th} and the $(t+d)^{th}$ positions (of the original n ordered samples) is:

$$\bar{x} = \frac{dN}{n+1}, \qquad (7)$$

with a variance:

$$\sigma_x^2 = \frac{Nd(n-d+1)(n+N+1)}{(n+1)^2\,(n+2)}. \qquad (8)$$

There are several interesting properties of the distribution (6). First, the probability that x of the new names fall in a given drawer depends only on the number of original samples defining the drawer, d+1, and not upon the index t which describes the position of the first name of the drawer in the original sample. In other words:

$$w_x(N,n;0,d) = w_x(N,n;1,d+1) = \ldots = w_x(N,n;t,t+d) = \ldots w_x(N,n;n+d-1,n+1). \qquad (9)$$

Secondly, we note that the averaged distribution is independent of the sampling distribution; roughly speaking, this is because on the average the new samples tend to "fill out" the sampling distribution in the same way that the original samples did.

---

* op.cit., pp. 58-63.

For only one new sample (N=1), the averaged distribution (6) reduces to the binomial distribution obtained from (1) by substituting the average probability, $\bar{p} = d/(n+1)$, between two ordered samples $\underline{d}$ units apart; however, for more than one new sample, the variance (8) is larger than that obtained by using the average parameter, $\bar{p}$, in the binomial distribution.

In the usual applications, both samples are large, with:

$$\lim_{\substack{N\to\infty \\ n\to\infty}} \frac{N}{N+n} = f \quad .$$

In this case:

$$w_x \simeq \binom{d-1+x}{d-1} f^x (1-f)^d \quad , \quad x= 0,1,2,..... \qquad (10)$$

a special case of the negative binomial distribution.

To answer the design problem posed at the beginning, we note that the averaged probability distribution is independent of the sampling distribution and of the original first or last name --- and thus depends only on the number of names to be stored in a given drawer. Hence, to equalize the probability of overflow among drawers of the same capacity, the same number of empty positions should be left in each drawer, no matter which portion of the alphabet is contained in the drawer.

Suppose that a given drawer begins with index $\underline{t}$, and the next drawer begins with $\underline{t+d}$; this scheme will preserve the original initial entry in each drawer, and there will be $\underline{d}$ names in each drawer to begin with. If the drawer has a capacity $\underline{c}$, then the overflow probability for this drawer is:

$$P_O = \sum_{x=c-d+1}^{N} w_x(N,n;0,d) \quad , \qquad (11)$$

which for the approximation (10) can be written as:

$$P_0 \simeq \binom{c}{d-1} f^{c-d+1} (1-f)^d \; {}_2F_1(c+1,1;\, c-d+2;\, f) \quad . \tag{12}$$

If the fraction f is small compared to the fraction of drawer space that is unused, the hypergeometric function ${}_2F_1$ may be approximated by unity.

If one has a system of r drawers, which begin with names of index $t_0 = 0, t_1, \ldots, t_{r-1}$ (original first entries are preserved in each drawer, except that zero always starts the first drawer), then an analysis similar to the one given above will give the joint distribution of x_0 falling in the first drawer, x_1 falling in the second drawer, $\ldots, x_{r-1}$ falling in the r^{th} drawer. We state without proof:

$$w_{x_0,x_1,\ldots,x_{r-1}}(N,n;t_0,t_1,\ldots,t_{r-1},t_r) = \frac{\prod_{k=0}^{r-1} \binom{t_{k+1}-t_k-1+x_k}{x_k}}{\binom{N+n}{n}} \tag{13}$$

with $\quad 0 = t_0 \leqslant t_1 \leqslant \cdots \leqslant t_{r-1} \leqslant t_r = N+1 \; ,$

and $\quad x_0 + x_1 + \cdots + x_{r-1} = N \; .$

In order to calculate the probability that <u>at least one</u> of the r drawers has overflowed, we must take the tail sum of (13).